PATENT COOPERATION TREATY

PCT

INTERNATIONAL PRELIMINARY EXAMINATION REPORT

(PCT Article 36 and Rule 70)

Applicant's or agent's file reference P27696WO RU	**FOR FURTHER ACTION** See Notification of Transmittal of International Preliminary Examination Report (Form PCT/IPEA/416)	
International application No. PCT/EP2004/000104	International filing date *(day/month/year)* 09.01.2004	Priority date *(day/month/year)* 24.01.2003
International Patent Classification (IPC) or both national classification and IPC G10L21/02, G10L11/02, G10L15/24		
Applicant SONY ERICSSON MOBILE COMMUNICATIONS AB ET AL.		

1. This international preliminary examination report has been prepared by this International Preliminary Examining Authority and is transmitted to the applicant according to Article 36.

2. This REPORT consists of a total of 6 sheets, including this cover sheet.

☒ This report is also accompanied by ANNEXES, i.e. sheets of the description, claims and/or drawings which have been amended and are the basis for this report and/or sheets containing rectifications made before this Authority (see Rule 70.16 and Section 607 of the Administrative Instructions under the PCT).

These annexes consist of a total of 5 sheets.

3. This report contains indications relating to the following items:

I ☒ Basis of the opinion
II ☐ Priority
III ☐ Non-establishment of opinion with regard to novelty, inventive step and industrial applicability
IV ☐ Lack of unity of invention
V ☒ Reasoned statement under Rule 66.2(a)(ii) with regard to novelty, inventive step or industrial applicability; citations and explanations supporting such statement
VI ☐ Certain documents cited
VII ☐ Certain defects in the international application
VIII ☐ Certain observations on the international application

Date of submission of the demand 05.08.2004	Date of completion of this report 14.03.2005
Name and mailing address of the international preliminary examining authority: European Patent Office - P.B. 5818 Patentlaan 2 NL-2280 HV Rijswijk - Pays Bas Tel. +31 70 340 - 2040 Tx: 31 651 epo nl Fax: +31 70 340 - 3016	Authorized Officer Quélavoine, R Telephone No. +31 70 340-3946

I. Basis of the report

1. With regard to the **elements** of the international application *(Replacement sheets which have been furnished to the receiving Office in response to an invitation under Article 14 are referred to in this report as "originally filed" and are not annexed to this report since they do not contain amendments (Rules 70.16 and 70.17))*:

Description, Pages

1-22	as originally filed

Claims, Numbers

1-14	received on 05.08.2004 with letter of 05.08.2004

Drawings, Sheets

1/7-7/7	as originally filed

2. With regard to the **language**, all the elements marked above were available or furnished to this Authority in the language in which the international application was filed, unless otherwise indicated under this item.

These elements were available or furnished to this Authority in the following language: , which is:

□ the language of a translation furnished for the purposes of the international search (under Rule 23.1(b)).

□ the language of publication of the international application (under Rule 48.3(b)).

□ the language of a translation furnished for the purposes of international preliminary examination (under Rule 55.2 and/or 55.3).

3. With regard to any **nucleotide and/or amino acid sequence** disclosed in the international application, the international preliminary examination was carried out on the basis of the sequence listing:

□ contained in the international application in written form.

□ filed together with the international application in computer readable form.

□ furnished subsequently to this Authority in written form.

□ furnished subsequently to this Authority in computer readable form.

□ The statement that the subsequently furnished written sequence listing does not go beyond the disclosure in the international application as filed has been furnished.

□ The statement that the information recorded in computer readable form is identical to the written sequence listing has been furnished.

4. The amendments have resulted in the cancellation of:

□ the description, pages:

□ the claims, Nos.:

□ the drawings, sheets:

5. □ This report has been established as if (some of) the amendments had not been made, since they have been considered to go beyond the disclosure as filed (Rule 70.2(c)).

(Any replacement sheet containing such amendments must be referred to under item 1 and annexed to this report.)

6. Additional observations, if necessary:

V. Reasoned statement under Article 35(2) with regard to novelty, inventive step or industrial applicability; citations and explanations supporting such statement

1. Statement

Novelty (N)	Yes:	Claims	1-14
	No:	Claims	
Inventive step (IS)	Yes:	Claims	1-3,5-7,13-14
	No:	Claims	4,8-12
Industrial applicability (IA)	Yes:	Claims	1-14
	No:	Claims	

2. Citations and explanations

see separate sheet

Re Item V

1. Reference is made to the following documents:
   D1: WO 02/29784 A (CLARITY LLC ; ERTEN GAMZE (US)) 11 April 2002 (2002-04-11)
   D2: WO 02/084644 A (DEUTSCHE TELECOM AG) 24 October 2002 (2002-10-24)

2. The document D1 is regarded as being the closest prior art to the subject-matter of claim 1, and shows (abstract, fig. 11,13-15) an audio-visual speech processing system in which the audio and video signals are analysed in parallel, and information from the video signal is used for speech detection and the selection of filters for noise removal in the audio signal.

   The subject-matter of claim 1 differs from this known system in that it also provides a multi-channel acoustic cancellation unit being specially adapted to perform near-end speaker detection and double-talk detection.

   The subject-matter of claim 1 is therefore new (Article 33(2) PCT).

   The problem to be solved by the present invention may be regarded as how to perform near-end speaker detection and double talk detection in an audio-visual interface.

   The solution to this problem proposed in claim 1 of the present application is considered as involving an inventive step (Article 33(3) PCT) because it is not obvious that the skilled person would use both audio and visual features to perform these detection. The skilled person would more likely use audio only.

   Claims 2, 3 are dependent on claim 1 and as such also meet the requirements of the PCT with respect to novelty and inventive step. Claim 13 and 14 are claims corresponding to the use of the system of claim 1 and as such also meet the requirements of the PCT with respect to novelty and inventive step.

3. The present application does not meet the criteria of Article 33(1) PCT, because the subject-matter of claim 4 does not involve an inventive step in the sense of Article 33(3) PCT.

3.1 The same document D1 is regarded as being the closest prior art to the subject-matter of claim 4 (see above disclosure an passages).
The subject-matter of independent claim 4 differs from the disclosure of D1 in that the noise reduction algorithm is a spectral subtraction method using the noise signal estimated during speech pauses while D1 uses filters dependent on the recognized visemes. This difference is only a simplification of the noise reduction system, coming back to a standard known method.

D2 (abstract) discloses this known noise estimation and spectral subtraction solution. The features disclosed in D1 and D2 would therefore be combined by the skilled person without exercise of any inventive skills in order to solve the corresponding problem. The proposed solution in independent claim 4 thus cannot be considered inventive (Article 33(3) PCT).

3.2 Dependent claims 8-12 do not contain any features which, in combination with the features of any claim to which they refer, meet the requirements of the PCT in respect of inventive step, see documents D1-2 and the corresponding passages cited in the search report.

3.3 The combination of the features of dependent claims 5-7 is neither known from, nor rendered obvious by, the available prior art. However, this combination results in the same subject-matter as defined by independent claim 1.

Re Item VII.

1. Contrary to the requirements of Rule 5.1(a)(ii) PCT, the relevant background art disclosed in the documents D1-5 is not mentioned in the description, nor are these documents identified therein.

Re Item VIII.

1. Claim 1 does not meet the requirements of Article 6 PCT in that the matter for which protection is sought is not clearly defined. The following functional statements do not enable the skilled person to determine which technical features are necessary to perform the stated function: "perform a near-end speaker detection and double-talk detection algorithm based on ..."
   The claim attempts to define the subject-matter in terms of the result to be achieved, which merely amounts to a statement of the underlying problem, without providing the technical features necessary for achieving this result.

PCT/EP2004/000104
SONY ERICSSON MOBILE COMMUNICATIONS AB
P27696WO

NEW CLAIMS

1. A noise reduction system with an audio-visual user interface, said system being specially adapted for running an application for combining visual features ($\underline{o}_{v,nT}$) extracted from a digital video sequence ($v(nT)$) showing the face of a speaker (S_i) with audio features ($\underline{o}_{a,nT}$) extracted from an analog audio sequence ($s(t)$), wherein said audio sequence ($s(t)$) can include noise in the environment of said speaker (S_i), said noise reduction system (200b/c) comprising

- means (101a, 106b) for detecting and analyzing said analog audio sequence ($s(t)$),
- means (101b') for detecting said video sequence ($v(nT)$), and
- means (104a+b, 104'+104'') for analyzing the detected video signal ($v(nT)$),

wherein a noise reduction circuit (106) of said noise reduction system is adapted to separate the speaker's voice from said background noise ($n'(t)$) based on a combination of derived speech characteristics ($\underline{o}_{av,nT} := [\underline{o}_{a,nT}{}^{T}, \underline{o}_{v,nT}{}^{T}]^{T}$) and outputting a speech activity indication signal ($\hat{s}_i(nT)$) which is obtained by a combination of speech activity estimates supplied by said analyzing means (106b, 104a+b, 104'+104''),

characterized by

a multi-channel acoustic echo cancellation unit (108) being specially adapted to perform a near-end speaker detection and double-talk detection algorithm based on acoustic-phonetic speech characteristics derived by said audio feature extraction and analyzing means (106b) and said visual feature extraction and analyzing means (104a+b, 104'+104'').

2. A noise reduction system according to claim 1,

characterized by

means (SW) for switching off an audio channel in case the actual level of said speech activity indication signal ($\hat{s}_i(nT)$) falls below a predefined threshold value.

3. A noise reduction system according to anyone of the claims 1 or 2,

characterized in that

said audio feature extraction and analyzing means (106b) is an amplitude detector.

4. A near-end speaker detection method reducing the noise level of a detected analog audio sequence ($s(t)$),

said method being characterized by the following steps:

- subjecting (S1) said analog audio sequence ($s(t)$) to an analog-to-digital conversion,
- calculating (S2) the corresponding discrete signal spectrum ($S(k\cdot\Delta f)$) of the analog-to-digital-converted audio sequence ($s(nT)$) by performing a Fast Fourier Transform (FFT),
- detecting (S3) the voice of said speaker (S_i) from said signal spectrum ($S(k\cdot\Delta f)$) by analyzing visual features ($\underline{o}_{v,nT}$) extracted from a simultaneously with the recording of the analog audio sequence ($s(t)$) recorded video sequence ($v(nT)$) tracking the current location of the speaker's face, lip movements and/or facial expressions of the speaker (S_i) in subsequent images,
- estimating (S4) the noise power density spectrum ($\Phi_{nn}(f)$) of the statistically distributed background noise ($\tilde{n}(t)$) based on the result of the speaker detection step (S3),
- subtracting (S5) a discretized version ($\tilde{\Phi}_{nn}(k\cdot\Delta f)$) of the estimated noise power density spectrum ($\tilde{\Phi}_{nn}(f)$) from the discrete signal spectrum ($S(k\cdot\Delta f)$) of the analog-to-digital-converted audio sequence ($s(nT)$), and
- calculating (S6) the corresponding discrete time-domain signal ($\hat{s}_i(nT)$) of the obtained difference signal by performing an Inverse Fast Fourier Transform (IFFT), thereby yielding a discrete version of the recognized speech signal.

5. A near-end speaker detection method according to claim 4,

characterized by the step of

conducting (S7) a multi-channel acoustic echo cancellation algorithm which models echo path impulse responses by means of adaptive finite impulse response (FIR) filters and subtracts echo signals from the analog audio sequence ($s(t)$) based on acoustic-phonetic speech characteristics derived by an algorithm for extracting visual features ($\underline{o}_{v,nT}$) from a video sequence ($v(nT)$) tracking the location of a speaker's face, lip movements and/or facial expressions of the speaker (S_i) in subsequent images.

6. A near-end speaker detection method according to claim 5, **characterized in** that said multi-channel acoustic echo cancellation algorithm performs a double-talk detection procedure.

7. A near-end speaker detection method according to anyone of the claims 4 to 6, **characterized in** that said acoustic-phonetic speech characteristics are based on the opening of a speaker's mouth as an estimate of the acoustic energy of articulated vowels or diphthongs, respectively, rapid movement of the speaker's lips as a hint to labial or labio-dental consonants, respectively, and other statistically detected phonetic characteristics of an association between position and movement of the lips and the voice and pronunciation of said speaker (S_i).

8. A near-end speaker detection method according to anyone of the claims 4 to 7, **characterized by** a learning procedure used for enhancing the step of detecting (S3) the voice of said speaker (S_i) from the discrete signal spectrum ($S(k\cdot\Delta f)$) of the analog-to-digital-converted version ($s(nT)$) of an analog audio sequence ($s(t)$) by analyzing visual features ($\underline{o}_{v,nT}$) extracted from a simultaneously with the recording of the analog audio sequence ($s(t)$) recorded video sequence ($v(nT)$) tracking the current location of the speaker's face, lip movements and/or facial expressions of the speaker (S_i) in subsequent images.

9. A near-end speaker detection method according to anyone of the claims 4 to 8, **characterized by** the step of correlating (S8a) the discrete signal spectrum ($S_\tau(k\cdot\Delta f)$) of a delayed version ($s(nT-\tau)$) of the analog-to-digital-converted audio signal ($s(nT)$) with an audio speech activity estimate obtained by an amplitude detection (S8b) of the band-pass-filtered discrete signal spectrum ($S(k\cdot\Delta f)$), thereby yielding an estimate ($\tilde{S}_i(f)$) for the frequency spectrum ($S_i(f)$) corre-

sponding to the signal ($s_i(t)$) which represents said speaker's voice as well as an estimate ($\tilde{\Phi}_{nn}(f)$) for the noise power density spectrum ($\Phi_{nn}(f)$) of the statistically distributed background noise ($n'(t)$).

10. A near-end speaker detection method according to claim 9,
characterized by the step of
correlating (S9) the discrete signal spectrum ($S_\tau(k\cdot\Delta f)$) of a delayed version ($s(nT-\tau)$) of the analog-to-digital-converted audio signal ($s(nT)$) with a visual speech activity estimate taken from a visual feature vector ($\underline{o}_{v,t}$) supplied by the visual feature extraction and analyzing means (104a+b, 104'+104''), thereby yielding a further estimate ($\tilde{S}_i'(f)$) for updating the estimate ($\tilde{S}_i(f)$) for the frequency spectrum ($S_i(f)$) corresponding to the signal ($s_i(t)$) which represents said speaker's voice as well as a further estimate ($\tilde{\Phi}_{nn}'(f)$) for updating the estimate ($\tilde{\Phi}_{nn}(f)$) for the noise power density spectrum ($\Phi_{nn}(f)$) of the statistically distributed background noise ($n'(t)$).

11. A near-end speaker detection method according anyone of the claims 9 or 10,
characterized by the step of
adjusting (S10) the cut-off frequencies of a band-pass filter (204) used for filtering the discrete signal spectrum ($S(k\cdot\Delta f)$) of the analog-to-digital-converted audio signal ($s(t)$) dependent on the bandwidth of the estimated speech signal spectrum ($\tilde{S}_i(f)$).

12. A near-end speaker detection method according to anyone of the claims 4 to 8,
characterized by the steps of

- adding (S11a) an audio speech activity estimate obtained by an amplitude detection of the band-pass-filtered discrete signal spectrum ($S(k\cdot\Delta f)$) of the analog-to-digital-converted audio signal ($s(t)$) to a visual speech activity estimate taken from a visual feature vector ($\underline{o}_{v,t}$) supplied by said visual feature extraction and analyzing means (104a+b, 104'+104''), thereby yielding an audio-visual speech activity estimate,
- correlating (S11b) the discrete signal spectrum ($S(k\cdot\Delta f)$) with the audio-visual speech activity estimate, thereby yielding an estimate ($\tilde{S}_i(f)$) for the frequency spectrum ($S_i(f)$)

corresponding to the signal ($s_i(t)$) which represents said speaker's voice as well as an estimate ($\tilde{\Phi}_{nn}(f)$) for the noise power density spectrum ($\Phi_{nn}(f)$) of the statistically distributed background noise ($n'(t)$) and

- adjusting (S11c) the cut-off frequencies of a band-pass filter (204) used for filtering the discrete signal spectrum ($S(k\cdot\Delta f)$) of the analog-to-digital-converted audio signal ($s(t)$) dependent on the bandwidth of the estimated speech signal spectrum ($\tilde{S}_i(f)$).

13. Use of a noise reduction system (200b/c) according to anyone of the claims 1 to 3 and a near-end speaker detection method according to anyone of the claims 5 to 13 for a video-telephony based application in a telecommunication system running on a video-enabled phone with a built-in video camera (101b') pointing at the face of a speaker (S_i) participating in a video telephony session.

14. A telecommunication device equipped with an audio-visual user interface,
characterized by
noise reduction system (200b/c) according to anyone of the claims 1 to 3.